CS4319 – Data Mining & Warehouses
Midterm 2

Name: Nhan Nguyen
Date: 04-24-17

1. The method where you reserve 2/3 for training and 1/3 for testing is (5 points)
   a. Holdout
   b. Cross Validation
   c. Bootstrap
   d. Stratified Training

2. Given two models of classification: (10 points)

   – Model M1: accuracy = 85%, tested on 30 instances

   – Model M2: accuracy = 75%, tested on 5000 instances

   What test would help to find which model is better?

   a. Test of Accuracy
   b. Test of Reliability
   c. Test of Significance
   d. Test of Comparability

3. K-means is (5 points)
   a) Medoid-based Hierarchical clustering
   b) Medoid-based Partitional clustering approach
   c) Centroid-based Hierarchical clustering
   d) Centroid -based Partitional clustering approach

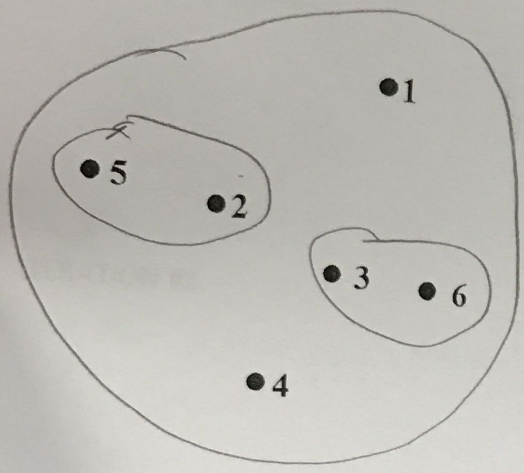
4. _____ measures how closely related are objects in a cluster (10 points)

   _____ measures how distinct or well-separated a cluster is from other clusters

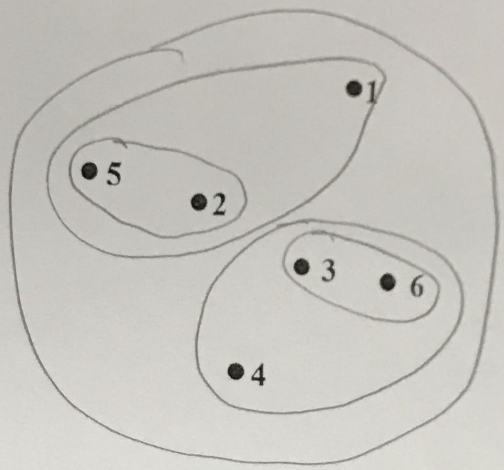
   a. Cluster Cohesion - Cluster Separation
   b. Cluster Separation - Cluster Cohesion
   c. Cluster Similarity - Cluster Distance
   d. Cluster Distance – Cluster Similarity

UHD
Univerity of Houston-Downtown

5. Show the order of merging by drawing a circle around them and numbering the order from 1,2... (15 points)



If MIN Measure used



If MAX Measure used

6. The method that predicts a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency is : (5 points)
   a. Correlation
   b. Regression
   c. Cohesion
   d. Separation

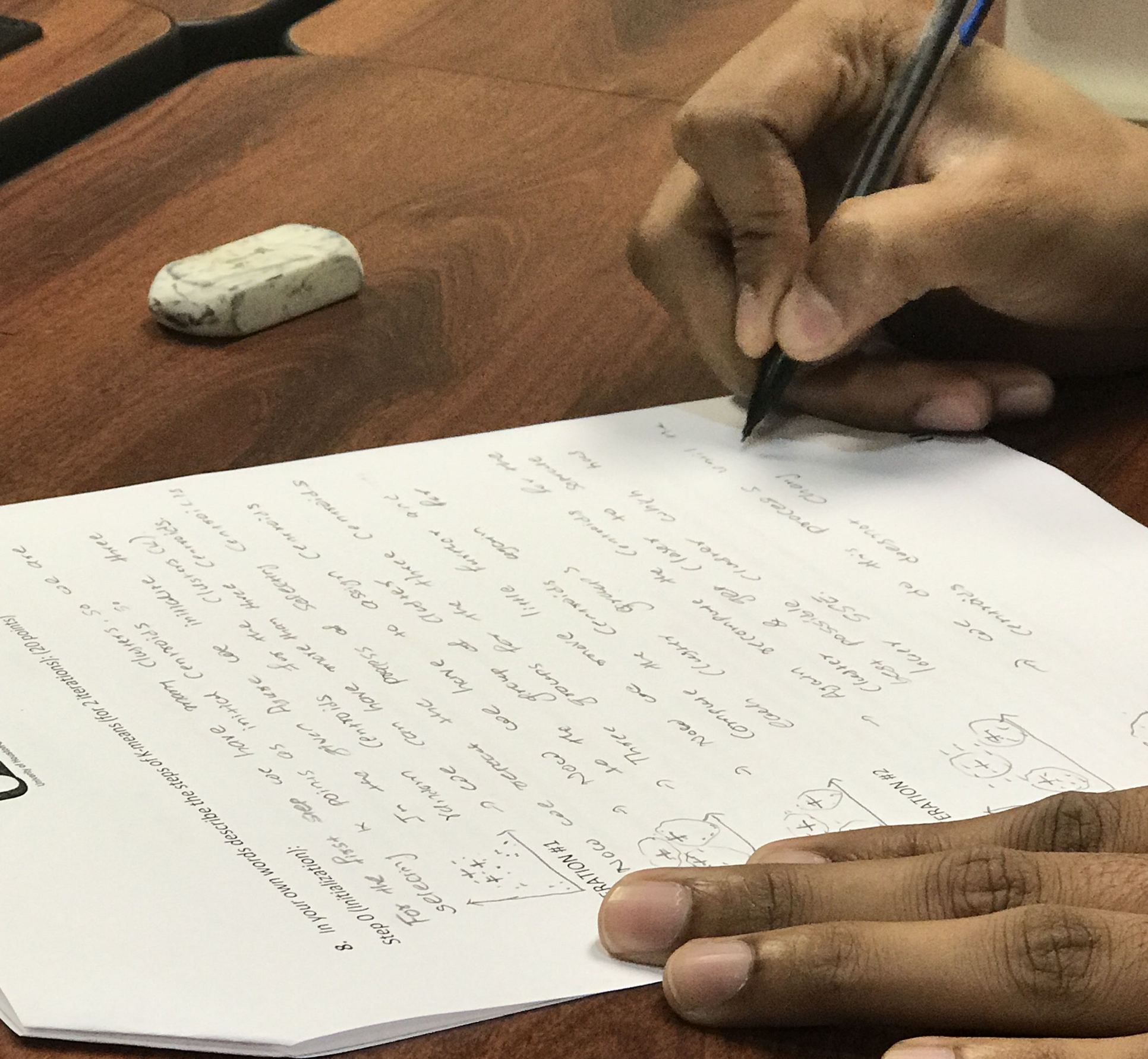
7. In your own words describe what Objective Function is and give an example? (20 points)

Types of Clusters: Objective Function

- Clusters Defined by an Objective Function
  - Finds clusters that minimize or maximize an objective function.
  - Enumerate all possible ways of dividing the points into clusters and evaluate the `goodness' of each potential set of clusters by using the given objective function. (NP Hard)
  - Can have global or local objectives.
    - Hierarchical clustering algorithms typically have local objectives
    - Partitional algorithms typically have global objectives
  - A variation of the global objective function approach is to fit the data to a parameterized model.
    - Parameters for the model are determined from the data.
    - Mixture models assume that the data is a `mixture' of a number of statistical distributions.

Types of Clusters: Objective Function ...

- Map the clustering problem to a different domain and solve a related problem in that domain
  - Proximity matrix defines a weighted graph, where the nodes are the points being clustered, and the weighted edges represent the proximities between points
  - Clustering is equivalent to breaking the graph into connected components, one for each cluster.
  - Want to minimize the edge weight between clusters and maximize the edge weight within clusters



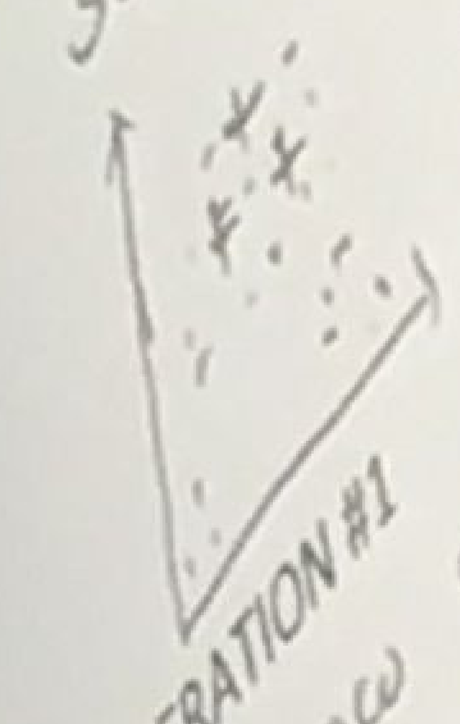
8. In your own words describe the steps of K-means (for 2 iterations): (20 points)

Step 0 (Initialization):

For the first step we have many clusters, so we are selecting k points as initial centroids.

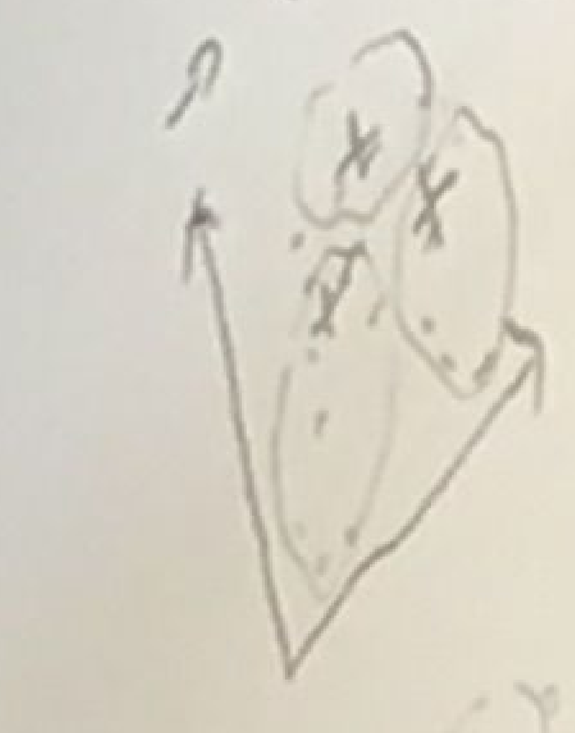
In the given there we initialize three random centroids for the clusters (k)

→ we can have more than three centroids



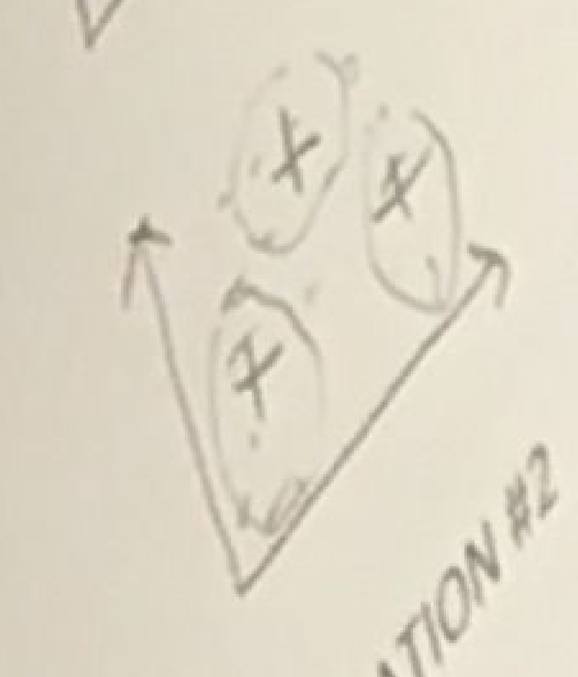
ITERATION #1

→ Now we repeat the process of selecting centroids



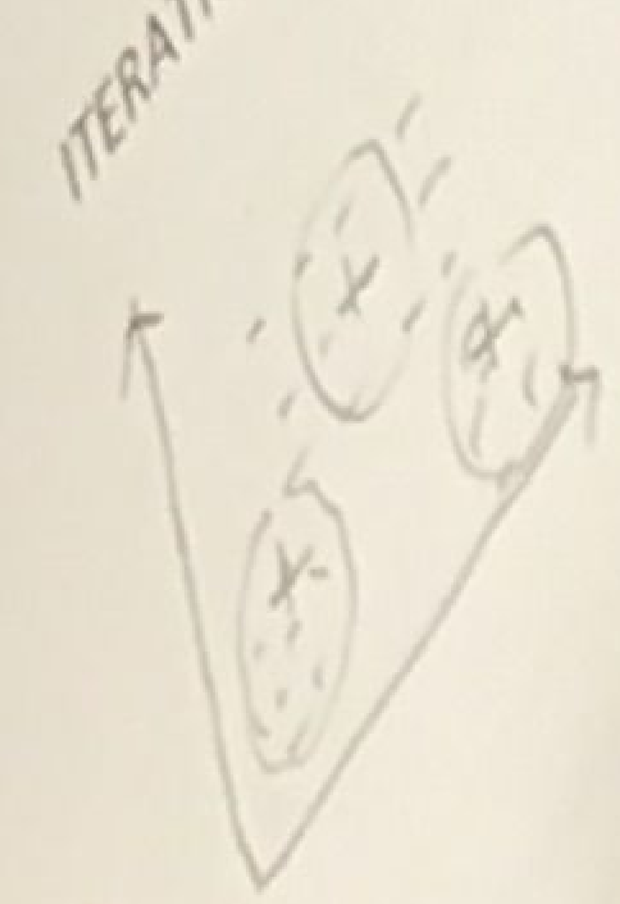
→ Now we have to assign centroids so the group of cluster

→ Three groups for the three centroids



→ Now we move little further and compute the centroids again for each cluster group's

ITERATION #2



→ Again recompute the centroids for the cluster & get closer to separate best possible cluster which has lower SSE.